Dispersion measures for ungrouped (raw) univariate data

BEA140 Quantitative Methods - Module 2

UNIVERSITY of TASMANIA

Dispersion

In these slides we will look at a number of dispersion measures for ungrouped (raw) univariate data.

In statistics **dispersion measures** attempt to give us an idea of how *stretched* or *squeezed* data points are.

The **range** of a data set is the difference between the maximum value and the minimum value.

$$\text{I.e. range} = X_{max} - X_{min}$$

Example: Going back to our ungrouped (raw) travel time data:

15	29	8	42	35	21	18	42	26

$$\text{range} = X_{max} - X_{min} = 42 - 8 = 34.$$

The range can be obtained in Excel using the `MAX` function minus the `MIN` function.

Dispersion - Variance (Defintion)

The **variance** of a data set is 'the mean/average of the values obtained from squaring the difference between each data point and the mean'.

$$\text{I.e.} \quad \sigma^2 = \frac{\Sigma(X_i - \mu_X)^2}{N} \quad \text{(population)}$$

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1} \quad \text{(sample)}$$

Note: More mathematically inclined students may wish to note that the reason the sample variance formula uses $n-1$ in the denominator (instead of n) is to correct a mathematically inherent bias as an estimation of the population variance.

Dispersion - Variance (Computation)

Although we won't, it is possible to show that the population and sample variance can be calculated using the following computationally friendly formulas:

$$\sigma^2 = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}}{N} \qquad \text{(population)}$$

$$s^2 = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1} \qquad \text{(sample)}$$

The population and sample variance can be obtained in Excel using the `VAR.P` and `VAR.S` functions respectively.

Note: When calculating variance (or standard deviation - see the next slide), it is more efficient and less error prone to use a table.

Dispersion - Variance Example

Going back to our ungrouped (raw) travel time data:

X_i	15	29	8	42	35	21	18	42	26
X_i^2	225	841	64	1764	1225	441	324	1764	676

$$\Sigma X_i = 15 + \ldots + 26 = 236$$

$$\Sigma X_i^2 = 225 + \ldots + 676 = 7324$$

$$s^2 = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1} = \frac{7324 - \frac{(236)^2}{9}}{8} = 141.94 \text{ (to 2 dp).}$$

Dispersion - Standard Deviation

(population) $\sigma = \sqrt{\sigma^2}$ (sample) $s = \sqrt{s^2}$

The population and sample standard deviation can be obtained in Excel using the `STDEV.P` and `STDEV.S` functions respectively.

Note: The standard deviation is by far the most commonly used measure of variation/dispersion.

Sanity Check: A "*rule of thumb*" is that the range is usually somewhere between 3 and 8 times the standard deviation.

I.e. for a population we usually have:

$$3\sigma \leq \text{range} \leq 8\sigma.$$

Dispersion - Standard Deviation Example

For our ungrouped (raw) travel time data, we obtained the variance $s^2 = 141.94$.

Hence the standard deviation is $s = \sqrt{141.94} = 11.91$ (to 2 dp).

Sanity Check: the range is $\frac{34}{11.91} \approx 2.85$ times the standard deviation, and hence outside the 3-8 band for our sanity check with this data!

Dispersion - The Empirical Rule

Interpreting the standard deviation is aided by what is famously referred to as *the empirical rule*, which states:

> *"For data sets that are normally distributed (normal distribution is introduced later) and as a 'rule of thumb' for any data set:*
>
> *(i) around* 68% *of the data will fall within one standard deviation of the mean;*
>
> *(ii) around* 95% *of the data will fall within two standard deviations of the mean; and*
>
> *(iii) around* 99% *of the data will fall within three standard deviations of the mean."*

Dispersion - Empirical Rule Example

With our ungrouped (raw) travel time data:

(i) the mean is $\overline{X} = 26.22$ (to 2 dp); and

(ii) the standard deviation is $s = 11.91$ (to 2 dp).

Hence the emperical rule suggests that (as a rule of thumb):

(i) 68% of the data will fall between 14.31 minutes and 38.13 minutes; and

(ii) 95% of the data will fall between 2.4 minutes and 50.04 minutes.

The **standard score** (z) of a single observation from a data set is the number of standard deviations that it is away from the mean.

$$\text{I.e. } z = \frac{X_i - \mu}{\sigma} \qquad \text{(population)} \qquad \text{and}$$

$$z = \frac{X_i - \overline{X}}{s} \qquad \text{(sample).}$$

With our ungrouped (raw) travel time data:

(i) for the mean we have $\overline{X} = 26.22$ (to 2 dp); and

(ii) for the standard deviation we have $s = 11.91$ (to 2 dp).

Hence the standard score of the observation 8 is:

$$z = \frac{8 - 26.22}{11.91} = -1.53 \text{ (to 2 dp).}$$

I.e. The observation 8 is -1.53 standard deviations below the mean.

. . . that's it for now, thanks for watching!

Don't forget that you can ask questions via:

(i) face-to-face lectures;

(ii) workshops or tutorials;

(iii) consultation hours; or

(iv) email.